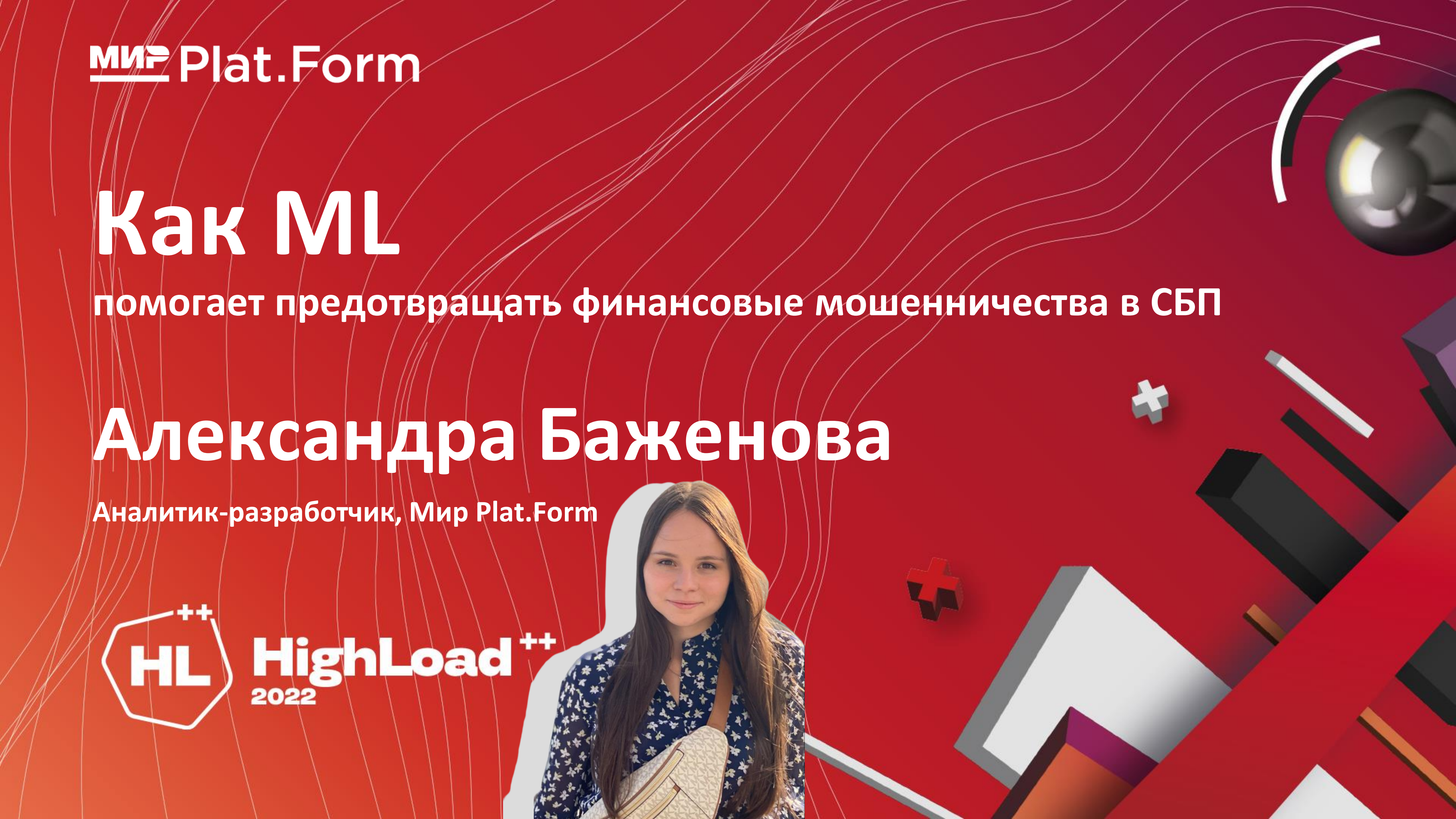

МИР Plat.Form
Как ML
помогает предотвращать финансовые мошенничества в СБП
Александра Баженова
Аналитик-разработчик, Мир Plat.Form
HL
HighLoad++
2022



Прив … ебя!
Три цифры номера карты
своей сообщу я тебе: 345
на ней написано!

Fraud Prevention



Какие типы fraud prevention-систем используют?



Rule based-системы

- Список правил, ограничивающий операции по одному или нескольким признакам
- Черные / серые списки
- Оценка рискованности операции

✔ **Хорошо интерпретируемые предсказания**

✘ **Необходима постоянная доработка системы**

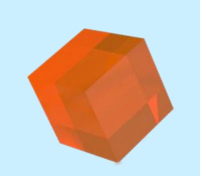


Machine learning-системы

- Вероятность мошеннической операции
- Оценка рискованности получателей
- Категоризация ТСП
- и другое

✘ **Предсказания не всегда интерпретируемы**

✔ **Реализуем процесс дообучения**



Machine Learning



Использование ML в задачах FP

–

Задача бинарной классификации

(при поиске фрода)

Полный цикл разработки системы предотвращения мошенничества

Разработка SML-системы с «нуля»

1. Подготовка данных
2. Расчет вторичных признаков
3. Индекс GINI
4. Матрица корреляции
5. Нормализация данных
6. Несбалансирован-ность выборки
7. Разработка модели ML
8. Оценка качества обучения модели

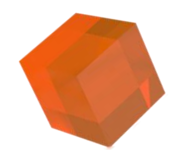


А как же машинное обучение без учителя?

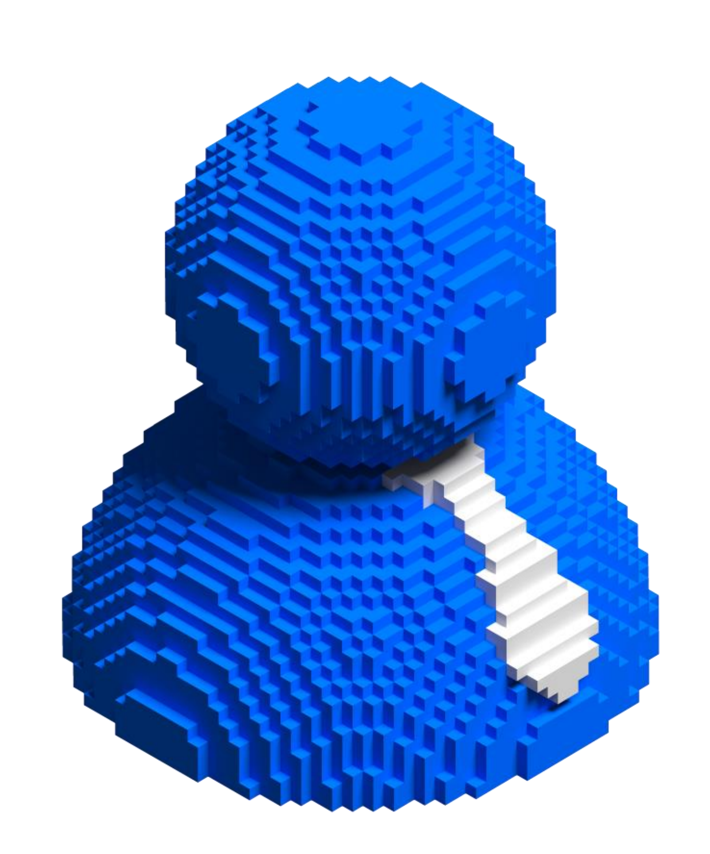
Варианты использования USML-систем в FP

Графовые методы

- Графы торгово-сервисных предприятий
- «Теория 3-х рукопожатий» ☻











Кластеризация

Кластеризация с целью объединения по близости финансовой активности

Социальная инженерия

за 2021 год количество
онлайн-платежей выросло на ~50%**

~33 млн платежей
Ноябрь 2020

~48 млн платежей
Ноябрь 2021

~56 млн платежей
Декабрь 2021



**по данным ЦБ в 2021 году*

***Ecom-операции ПС «Мир»*

Система Быстрых Платежей

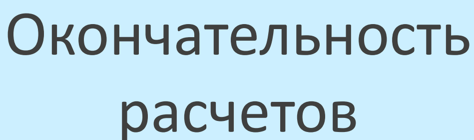




на проведение операции
на практике еще меньше



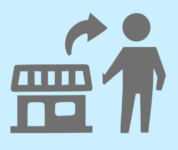


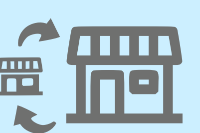

Окончательность
расчетов

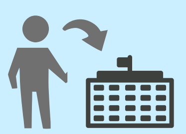








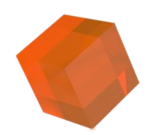






C2C C2B B2C B2B C2G





Отличия данных

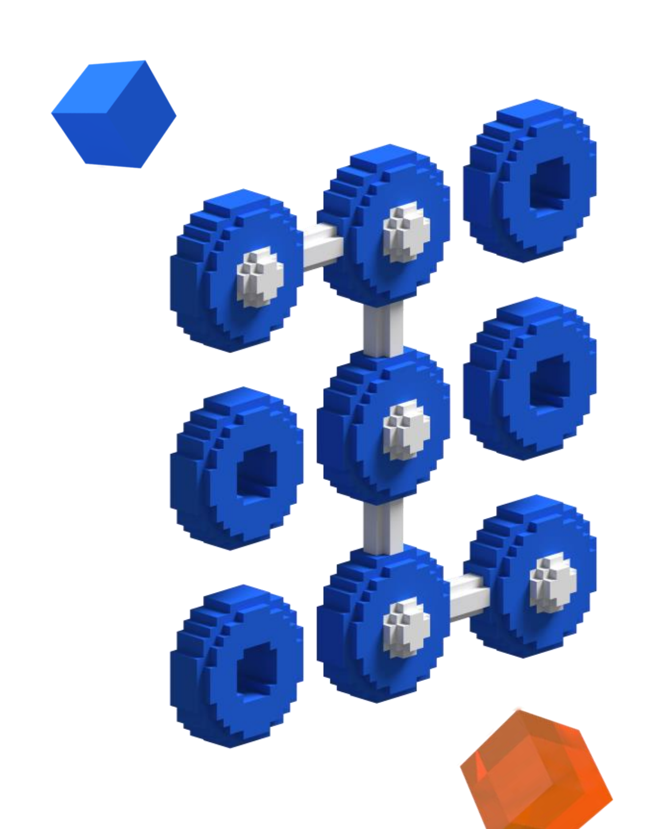


		НСПК Платежные системы	Банки
Данные для построения агрегат	Персональные данные Отправителя	✗	✓
Разметка	Обратная связь по операциям	✗	✓
Данные для обучения и построения агрегат	История операций Отправителя	✓	✓
Данные для обучения и построения агрегат	История операций Получателя	✓	✓ ✗

Машинное обучение без разметки

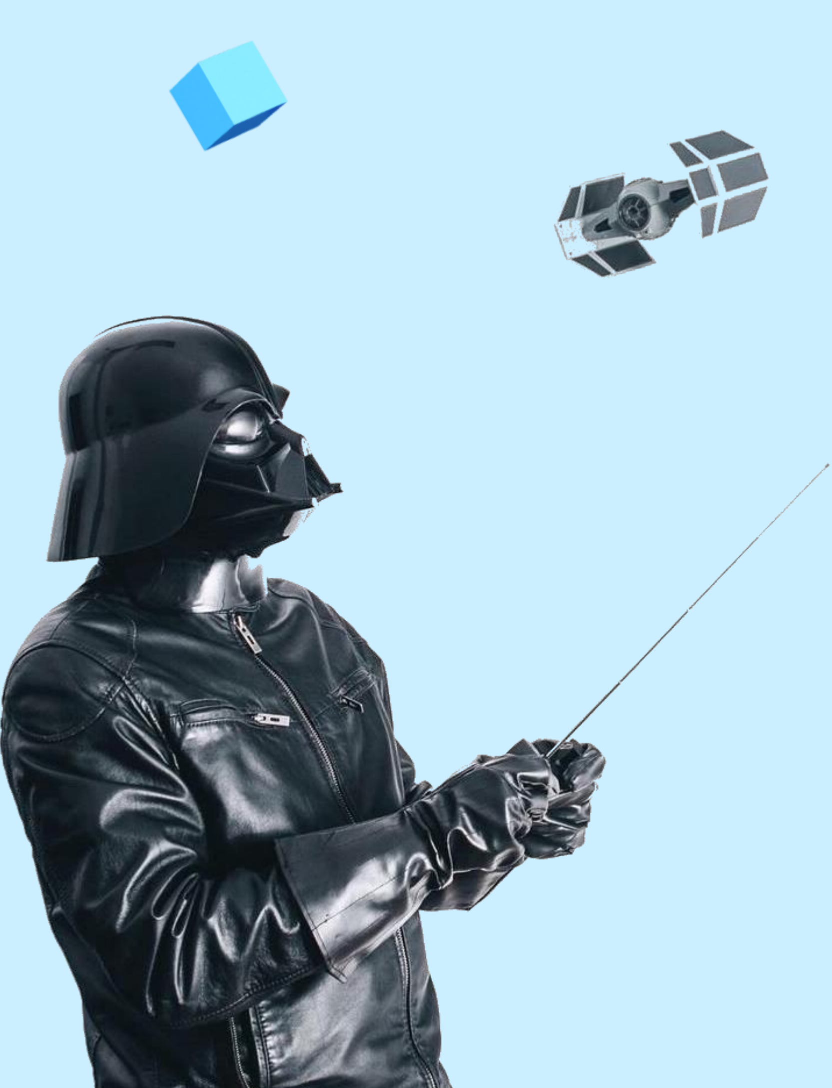
Поиск аномалий





Поиск аномалий	Мошеннические операции не выявлены
Алгоритмы кластеризации	Выявлено нецелевое использование системы
Нейронная сеть	Выявлены переводы определенного типа, не являющиеся мошенничеством

Симулятор мошенника



Инструмент обогащения транзакционных данных синтетическими мошенническими транзакциями в соответствии с условиями генерации для последующего обучения и тестирования алгоритмов ML



Генерирует синтетические транзакции



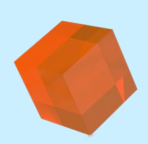
Сохраняет структуру данных СБП



Использует реальные данные для расчета статистических признаков генерации



Имитирует максимально приближенное поведение злоумышленников



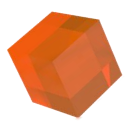
Симулятор мошенника



Компоненты симулятора/роли



«Мошенник»



«Клиент-Пострадавший»

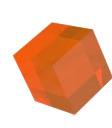


«Клиент-Знакомый»

Симулятор мошенника

Вывод средств с мошеннических счетов

В симуляторе заложена вероятность вывода средств через СБП,
а также параметризирована сумма вывода по отношению к хищению



Зачем?

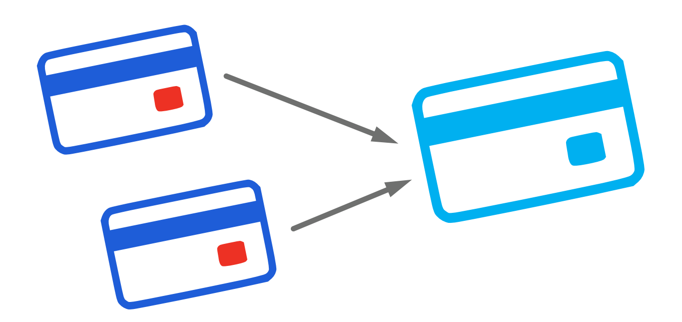
Мошенники стремятся в кратчайшие сроки обналичить деньги !

Как? Подставные карты!

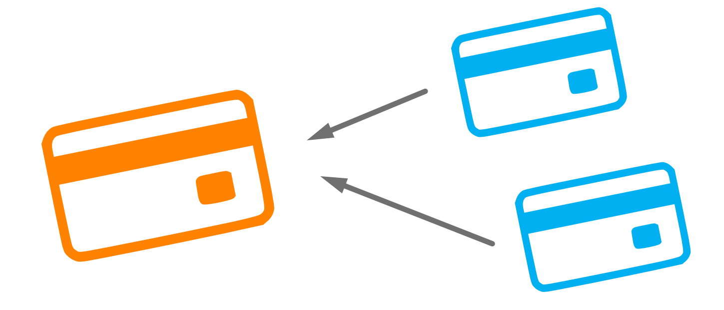
Для СБП номера телефонов



Мошенник использует несколько подставных карт



Группа мошенников использует подставную карту



Возможен частичный вывод средств
через разные каналы

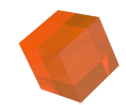
Возможен вывод средств через другие каналы,
не через СБП



Симулятор мошенника







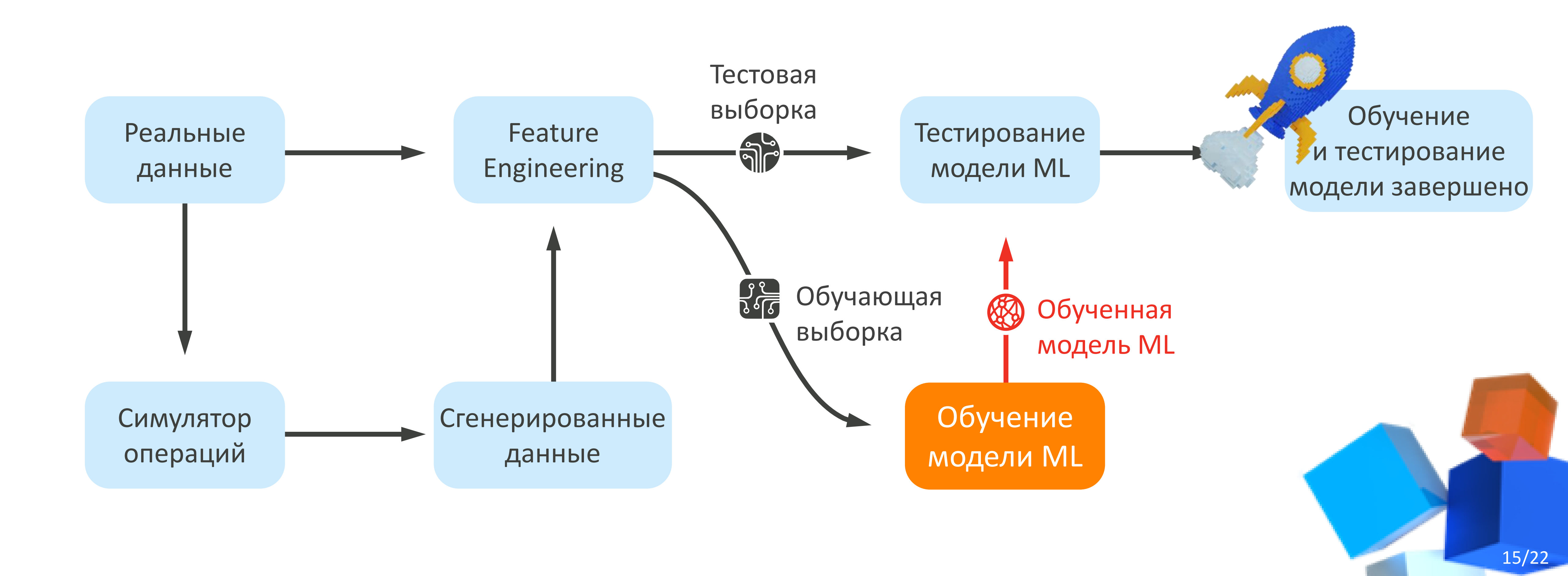
Основные параметры

- Доля транзакций по типам
- Вероятность вывода средств
- Максимальное количество жертв
- Период
- Максимальная доля мошеннических транзакций

	Основные требования к сценарию	Опциональные требования, задаваемые в параметрах
 Сценарии **С ОДНИМ МОШЕННИКОМ**	• Жертва не знает мошенника • Возможный полный или частичный вывод средств	• Одинаковые суммы • Частота мошенничеств • Регистрация клиента в СБП • Количество фрода у мошенника
 Сценарии **С НЕСКОЛЬКИМИ МОШЕННИКАМИ**	• Резкая активность • Клиенты-получатели знакомы • Возможный полный или частичный вывод средств	• Ограниченный период проведения мошеннических операций на группу клиентов-получателей

Процесс подготовки данных

Обучение и тестирование модели (1/2)



Процесс подготовки данных

Обучение и тестирование модели (2/2)



Подача реальных данных на вход симулятору мошеннических операций

Получение сгенерированных мошеннических и легальных операций

Объединение реальных данных и сгенерированных для подготовки данных для запуска модели ML

Подача обучающей выборки на вход модулю обучения модели ML

Получение обученной модели ML

Передача тестовой выборки для оценки обученной модели ML

Получение списка подозрительных операций и анализ результатов

Какие методы мы исследовали?

Unsupervised machine learning

- PageRank
- Elliptic Envelope
- Isolation Forest
- Local Outlier Factor
- K-means
- DBSCAN
- Autoencoder + GAN



Низкая эффективность выявления фрода

Supervised machine learning



- Logistic Regression
- Random Forest
- LightGBM
- CatBoost
- Neural Network



Результаты испытаний положительные

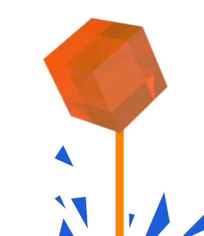
Результаты алгоритмов Supervised Machine Learning





Качество





Скорость



Результаты тестирования на данных симулятора

80% фрода найдено при 1 заблокированной легальной из 1000

Результаты тестирования на реальных данных *(июль 2020)*

54-60% фрода найдено при 5 заблокированных легальных из 1000

Оценка производительности алгоритмов SML *(скоринг транзакций)*

500 000 – 700 000 операций в секунду

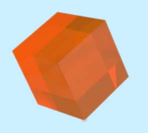
Экономическая эффективность

21%
денежных средств
сохраняет система FP
с использование ML

Стоимость владения



$$CO = \frac{CAPEX}{5} + OPEX$$



CO – cost of ownership
CAPEX – капительные затраты
OPEX – операционные затраты
Срок амортизации – 5 лет



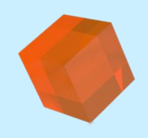


Экономический эффект



$$E = O * P * (1 - FAR)$$

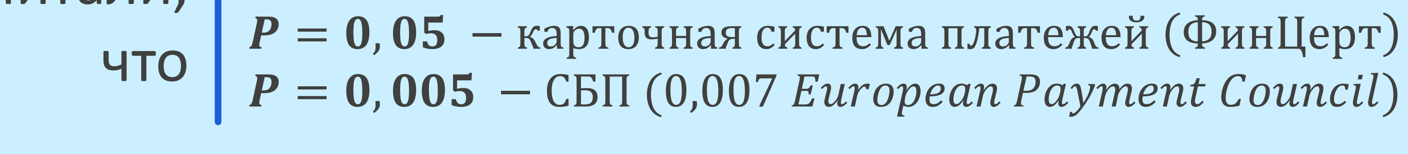
O – годовой оборот
P – вероятность возникновения мошенничества (уровень ВР)
FAR – false acceptance rate



Считали, что

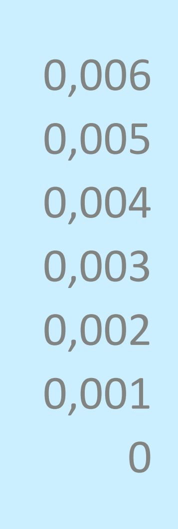


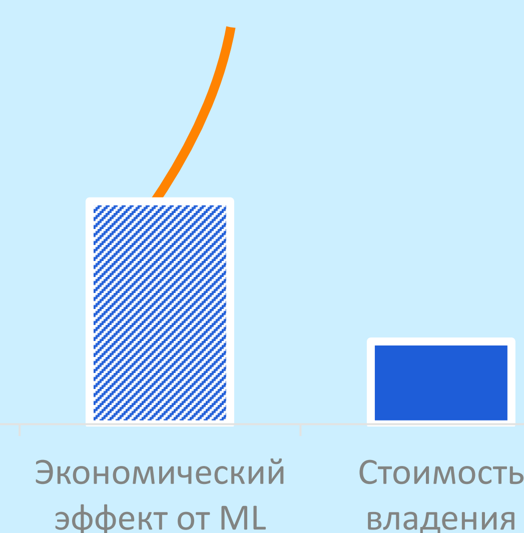
$$FAR = 0,3$$



$P = 0,05$ – карточная система платежей (ФинЦерт)
$P = 0,005$ – СБП (0,007 *European Payment Council*)

Доля годового оборота, проценты



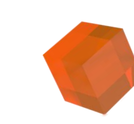




Результаты

Признаковое пространство

- Оконные признаки по суммам
- Оконные признаки по количествам
- Поведенческие признаки



Симулятор синтетических операций

- Реальные данные
- Мошеннические схемы
- Оценка генерации

Градиентный бустинг

- В реализации LightGBM

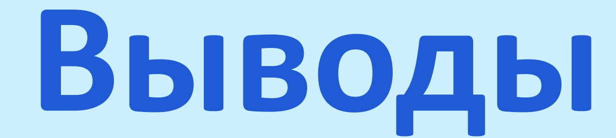


Качественные показатели

- FAR = 0,6
- FRR = 0,005
- 500 000 TPS

Необходимый объем исторических данных

- полгода







Методы SML способны эффективно выявлять мошеннические операции в СБП



Методы UML способны выявлять нецелевое использование СБП



Подтверждена возможность обучения моделей SML на синтетических транзакционных данных



Ожидается, что применение моделей SML обеспечит снижение уровня мошенничества

А ВЫ ИСПОЛЬЗУЕТЕ **ML?**





Александра Баженова
Аналитик-разработчик, Мир Plat.Form

